911 – Predicting Survival in Cardiac Arrest

Leveraging NEMSIS Data for Advanced Analysis and Improved Prediction of Emergency Response

Lillith Chute, Cristal Wilson Lobo, and Jonas Eichenlaub

Agenda

- Project Synopsis
- Data Preprocessing and Imputation
- Feature Selection
- Model
- Project Results
- Questions

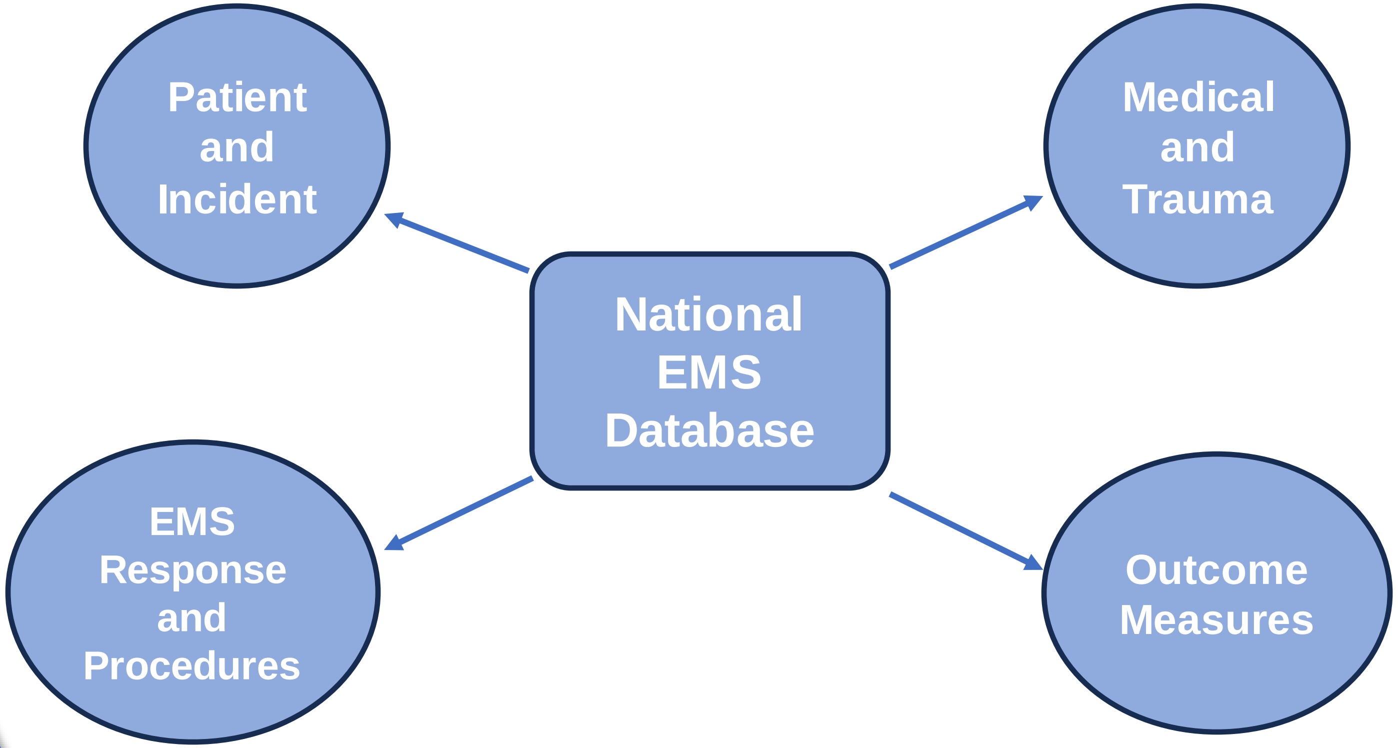
Project Synopsis

Problem Statement: Developing a methodology to predict outcomes in cardiac arrest cases by leveraging EMS data.

Goal: Utilizing the vast EMS data from NEMSIS to analyze patterns and build predictive models for cardiac arrest cases.

Challenge : Identifying target variable for outcome prediction , impute missing data and prepare dataset

Key Focus : Examining trends, response patterns, and patient outcomes within the EMS data.



Data Preprocessing

Source of Data:

- Source Folder : ProcessedDataCA.zip
- Types of Files :
  - Comprehensive Files: Contain multiple parameters related to the emergency medical system, regional data, and situational context.
  - Single Column Files (FACT*.csv): Detail specific events or occurrences in the emergency medical system.
  - PCR Key (Patient Care Report) - Acts as a common identifier across all files

Preprocessing Methodology:

- Primary Dataframe Creation:
  - Merge comprehensive files to form a primary merged dataframe. This serves as the foundation for further data appending and analysis.
- Appending FACT*.csv Files:
  - Iteratively process all FACT*.csv files. Append columns to the primary dataframe, ensuring no duplication. Verify the presence and consistency of values in these columns.
- Loading NEMSIS XSD Code-Value Pairs:
  - Extract and load code-value pairs for each element from NEMSIS XSD files. This step is crucial for understanding and translating coded data into interpretable information.
- Mapping Code-Value Pairs:
  - Systematically map the loaded code-value pairs to respective columns in the dataframe.

Challenges:

- Missing code value pairs
- Abundance of unknown values
- Identifying target outcome

Creation of Outcome Variable:

- Categorizing patient outcomes as a binary output (Alive/Dead) based on 'eArrest_18'.

Element Value	Outcome
Expired in ED	Dead
Expired in the Field	Dead
Ongoing Resuscitation in ED	Unknown
ROSC in the Field	Alive
ROSC in the ED	Alive
Ongoing Resuscitation by Other EMS	Unknown

Data Cleaning:

- Drop irrelevant date/time columns.
- Identifying and handling missing/mis-coded values with NA

Data Imputation:

- Distinguish between categorical and numeric features
    - Numeric variables – Replace missing values using median value
    - Categorical variables - Replace missing values using most frequent value

```
<bound method NDFrame.head of          PcrKey USCensusRegion    USCensusDivision  ...   eResponse_09      eArrest_17 eSituation_11
0        25944387                                  ...  None/No Delay  Not Applicable          S39.91
1        71121582        South      East South Central  ...  None/No Delay         9901003          I46.9
2        71122461        South      East South Central  ...  None/No Delay         9901047          I46.9
3        71122902        South      East South Central  ...  None/No Delay         9901035          I46.9
4        71123389        South      East South Central  ...  None/No Delay  Not Applicable          I46.9
...           ...          ...                    ...  ...            ...             ...           ...
448679  131801464      Midwest      East North Central  ...  None/No Delay         9901035            nan
448680  131801585      Midwest      East North Central  ...  None/No Delay         9901005            nan
448681  131801624      Midwest      East North Central  ...  None/No Delay    Not Recorded            nan
448682  131801707      Midwest      East North Central  ...   Not Recorded         9901067            nan
448683  131801811      Midwest      East North Central  ...  None/No Delay    Not Recorded            nan

[448684 rows x 126 columns]>
```

Data Imputation

Feature Selection

Lasso:
- Lasso Logistic Regression, selected features in best model
- Advantages: Flexible, weight on penalty term can be tuned
- Disadvantages: Suggests including many features, long runtime

Principal Component Analysis:
- Run PCA Logistic Regression with varied number of components
- Select most significant feature in top 20 components of best model
- Advantages: Selects most important feature from each dimension
- Disadvantages: Hard to interpret, can select already selected features

Univariate Analysis:
- Rank features, then select top 20
- Used Mutual Information & F Statistic as metrics
- Advantages: Easy to interpret, no underlying model assumptions
- Disadvantages: Ignores feature dependencies

Average of Algorithms
- Averaged features selected by Lasso, PCA, and Univariate, select top 20
- Advantages: Includes features selected by multiple approaches
- Disadvantages: Has no real statistical / methodological foundation

Subject Matter Expert (SME):
- Leveraged Theresa May's annotations and Brandon Skwarto's comments to judgmentally select 20 features
- Advantages: Based on medical expertise and human logic
- Disadvantages: Could exclude unintuitive features

Features Selected (balanced data)

Feature ID	Feature Name	SME	Avg	Lasso	PCA	Uni	Total
eArrest_02	Cardiac Arrest Etiology	X	X	X	X	X	5
eArrest_05	CPR care provided prior to EMS arrival	X	X	X	X	X	5
eArrest_01	Cardiac Arrest	X	X	X		X	4
eArrest_07	AED Use Prior to EMS Arrival	X	X	X		X	4
eArrest_11	First Monitored Arrest Rhythm of the Patient	X	X	X		X	4
ePatient_13	Gender	X	X	X	X		4
USCensusDivision	Census Division	X	X	X	X		4
ageinyear	Age in Years	X	X	X		X	4
EMSSceneTimeMin	EMS Scene Time	X	X	X		X	4
EMSTransportTimeMin	EMS Transport Time	X	X	X		X	4
eResponse_15	Level of Care of this Unit	X			X	X	3
eDisposition_16	EMS Transport Method		X	X	X	X	4
eScene_08	Triage Classification for MCI Patient		X	X	X	X	4
eDisposition_17	Transport Mode from Sence		X	X		X	3
eOutcome_02	Hospital Disposition		X	X		X	3
ePayment_01	Primary Method of Payment		X	X	X		3
ePayment_50	CMS Service Level		X	X		X	3
eProcedures_02	Procedure Performed Prior to EMS Care		X		X	X	3
eResponse_05	Type of Service Requested		X		X	X	3
NasemsoRegion	Region Name		X	X	X		3
EMSTotalCallTimeMin	EMS Total Call Time		X	X		X	3
eArrest_04	Arrest Witnessed By	X					1
eArrest_16	Reason CPR/Resuscitation Discontinued	X					1
ePatient_14	Patient Race	X					1
eResponse_10	Type of Scene Delay	X					1
eResponse_11	Type of Transport Delay	X					1
eVitals_26	Level of Responsiveness (AVPU)	X					1
EMSSystemResponseTimeMin	EMS System Response Time	X					1
eVitals_10	Heart Rate	X					1
eVitals_16	End Tidal Carbon Dioxide (ETCO2)	X					1
Urbanicity	Urbanicity	X					1

Models

Naïve Bayes:

- Simple and fast.
- Good for large datasets.
- Makes for a good baseline model

Random Forest:

- As ensemble method it's good against overfitting
- Effective with categorical and continuous data
- Capture non-linear relationships

XGBoost:

- Known for delivering high performance models.
- Handles various data types.
- Has regularization
- Computationally efficient and handles large datasets.

	Naive Bayes	Random Forest	XG Boost	Average Performance
Lasso	0.835	0.86	0.884	0.859666667
PCA	0.77	0.79	0.8	0.786666667
Univariate	0.84	0.86	0.87	0.856666667
Average of Algorithms	0.84	0.86	0.88	0.86
Subject Matter Expert	0.87	0.9	0.91	0.893333333

Model Performances (balanced data)

Project Results

- Subject-matter expert selected features performed the best.
- Lasso was within a couple of percenatage points.
- In all cases of feature selection XGBoost, as a model, was the best performer.
- Based on balanced data and features selected, the best model is 91% accurate at predicting CA survival.

Next Steps

- Try alternative data imputation techniques
- Use three outcome target variable: Dead, Alive, Coma
- Explore feature engineering
- Include medication & medical procedure features
- Judgmentally combine SME & algorithm selected features
- Experiment with different model scoring metrics (like ROC AUC)
- Explore other error analysis techniques

GitHub

https://github.com/ds5110/project-fall23-LillithChute/tree/main

QUESTIONS

Citations

Fonti, Valeria, and Eduard Belitser. "Feature selection using lasso." VU Amsterdam research paper in business analytics 30 (2017): 1-25.

Hua, Jianping, Waibhav D. Tembe, and Edward R. Dougherty. "Performance of Feature-Selection Methods in the Classification of High-Dimension Data." Pattern Recognition 42, no. 3 (March 1, 2009): 409–24. https://doi.org/10.1016/j.patcog.2008.08.001.

Odhiambo Omuya, Erick, George Onyango Okeyo, and Michael Waema Kimwele. "Feature Selection for Classification Using Principal Component Analysis and Information Gain." Expert Systems with Applications 174 (July 15, 2021): 114765. https://doi.org/10.1016/j.eswa.2021.114765.

Song, Fengxi, Zhongwei Guo, and Dayong Mei. "Feature Selection Using Principal Component Analysis." In Engineering Design and Manufacturing Informatization 2010 International Conference on System Science, 1:27–30, 2010. https://doi.org/10.1109/ICSEM.2010.14.